

Technical Report TR32, May 2004

Technical Report: OSU-CISRC-5/04-TR32

Department of Computer and Information Science

The Ohio State University, Columbus, OH 43210-1277, USA

Web site: <http://www.cis.ohio-state.edu/research/tech-report.html>

Ftp site: <ftp.cis.ohio-state.edu>

Login: **anonymous**

Directory: **pub/tech-report/2004**

File in pdf format: **TR32.pdf**

BINARY AND RATIO TIME-FREQUENCY MASKS FOR ROBUST SPEECH RECOGNITION

Soundararajan Srinivasan ^{a,1}, Nicoleta Roman ^b,
DeLiang Wang ^b

^a*Biomedical Engineering Center
The Ohio State University
Columbus, OH 43210, USA
srinivasan.36@osu.edu*

^b*Department of Computer and Information Science & Center for Cognitive
Science
The Ohio State University
Columbus, OH 43210, USA
niki,dwang@cis.ohio-state.edu*

Abstract

A time-varying Wiener filter extracts a speech signal from a mixture using the *a priori* signal-to-noise ratio in a local time-frequency unit. We estimate this ratio using a binaural processor and derive a ratio time-frequency mask. This mask is used to extract the speech, which is then fed to a conventional speech recognizer operating in the cepstral domain. We compare the performance of this system with a missing-data recognizer that operates in the spectral domain using the time-frequency units dominated by speech. For use by the missing-data recognizer, the same processor is used to estimate an ideal time-frequency binary mask, which selects the speech if it is stronger than the interference in a local time-frequency unit. We find that the performance of the missing-data recognizer is better on a small vocabulary recognition task but the performance of the conventional recognizer is substantially better when the vocabulary size is larger.

Key words: Ideal binary mask, Ratio mask, Robust speech recognition, Missing-data recognizer, Binaural processing, Speech segregation.

1 Introduction

The performance of automatic speech recognizers (ASRs) degrades rapidly in the presence of noise and other distortions (Gong, 1995; Lippmann, 1997). Speech recognizers are typically trained on clean speech and face a problem of mismatch when used in conditions where speech occurs simultaneously with other sound sources. To mitigate the effect of this mismatch on recognition, noisy speech is typically preprocessed by speech enhancement algorithms, such as microphone arrays (Brandstein and Ward, 2001; Cardoso, 1998; Ehlers and Schuster, 1997; Hughes et al., 1999), computational auditory scene analysis (CASA) systems (Rosenthal and Okuno, 1998; Wang and Brown, 1999) or spectral subtraction techniques (Boll, 1979; Droppo et al., 2002). Microphone arrays require the number of sensors to increase as the number of interfering sources increases. Monaural CASA systems employ harmonicity as the primary cue for grouping acoustic components corresponding to speech. These systems, however, do not perform in time-frequency regions that are dominated by unvoiced speech. Spectral subtraction systems typically assume stationary noise. Hence, in the presence of non-stationary noise sources, their performance is not adequate for recognition (Cooke et al., 2001). If samples of the corrupting noise source are available *a priori*, a model for the noise source can additionally be trained and noisy speech may be jointly decoded using the trained models of speech and noise (Gales and Young, 1996; Varga and Moore, 1990) or enhanced using linear filtering methods (Ephraim, 1992). However, in many realistic applications, adequate amounts of noise samples are not available *a priori* and hence training of a noise model is not feasible.

Recently a missing-data approach to speech recognition in noisy environments has been proposed by Cooke et al. (2001). This method is based on distinguishing between reliable and unreliable data. When speech is contaminated by additive noise, some time-frequency (T-F) units contain predominantly speech energy (reliable) and the rest are dominated by noise energy. The missing-data method treats the latter T-F units as missing or unreliable during recognition (see Section 4.2). Missing T-F units are typically identified using spectral subtraction. The performance of missing-data recognizer is significantly better than the performance of a system using spectral subtraction for speech enhancement followed by recognition of enhanced speech (Cooke et al., 2001).

A potential disadvantage of the missing-data recognizer is that recognition is performed in the spectral or T-F domain. It is well known that recognition using cepstral coefficients yields a superior performance compared to recognition using spectral coefficients under clean speech conditions (Davis and Mermel-

¹ Corresponding author. Tel.: 1-614-292-7402.

stein, 1980). The superiority of the cepstral features stems from the ability of the cepstral transformation to separate vocal-tract filtering from excitation source in speech production (Rabiner and Juang, 1999). Since the missing-data recognition is based on treating local T-F units as missing or unreliable T-F features during recognition, it is coupled with a spectral or T-F representation. Any global transformation of the spectral features (e.g. cepstral transformation) smears the information in the noisy T-F units, preventing its effective marginalization. Attempts to adapt the missing-data method to the cepstral domain have centered around reconstruction or imputation of the missing values in the spectral domain followed by transformation to the cepstral domain (Cooke et al., 2001; Raj et al., 2000). This reconstruction is typically based either on the speech recognizer itself or on other trained models of speech. The success of these model-based imputation techniques depend on the adequacy of reliable data for identification of the correct speech model for imputation. In addition, errors in imputation procedures affect the performance of the system even when the model is correctly identified.

Another potential drawback of the missing-data recognizer, which has not been well studied, is the problem of data paucity. The amount of “reliable” data available to the recognizer is a function of both SNR and the frequency characteristics of the noise source. A decrease in SNR, as well as an increase in the bandwidth of the noise source causes an increase in the amount of missing data. This leads to a deterioration in performance for a small vocabulary task (Cooke et al., 2001). The reduction in reliable data may pose an additional problem for recognition with larger vocabulary sizes. Paucity of reliable data constrains the missing-data recognizer to use only a small portion of the total T-F acoustic model space. This reduced space may be insufficient to differentiate between a large number of competing hypotheses during decoding. In this paper, we study this issue by comparing the performance of the missing-data recognizer on two tasks with different vocabulary sizes.

Binaural CASA systems that compute an ideal binary mask have been used as front-ends for the missing-data recognizer previously (Palomaki et al., 2004; Roman et al., 2003). A T-F unit in the ideal binary mask is labeled 1 or reliable if the corresponding T-F unit of the noisy speech contains more speech energy than interference energy; it is labeled 0 or unreliable otherwise. We employ a recent binaural speech segregation system (Roman et al., 2003) to estimate an ideal binary T-F mask. This mask is fed to the missing-data recognizer and recognition is performed in the spectral domain.

The minimum mean-square error (MMSE) based short-time spectral amplitude estimator, which utilizes *a priori* SNR in a local T-F unit, has been used previously to effectively enhance noisy speech (Ephraim and Malah, 1984). *a priori* SNR can be obtained if premixing speech and noise signals are available. Roman et al. (2003) have shown that in a narrow frequency band, there

exists a systematic relationship between *a priori* SNR and values of the binaural cues of interaural time differences (ITD) and interaural intensity differences (IID). Motivated by this observation, we estimate an ideal ratio T-F mask using statistics collected for ITD and IID at each individual frequency. A unit in the ratio mask is a measure of the speech energy to total energy (speech and noise) in the corresponding T-F unit of noisy signal. The ratio mask is then used to enhance the speech, enabling recognition using Mel-Frequency Cepstral Coefficients (MFCCs). We use “conventional recognizer” to refer to a continuous density hidden Markov model (HMM) based ASR using MFCCs as features.

We compare the performance of the conventional recognizer to that of the missing-data recognizer on a robust speech recognition task. In particular, we examine the effect of vocabulary size on the performance of the two recognizers. We find that on a small vocabulary task, the missing-data recognizer outperforms the conventional ASR. Our finding is consistent with a previous comparison using a binaural front-end made on a small vocabulary “cocktail-party” recognition task (Glotin et al., 1999; Tessier et al., 1999). The accuracy of results obtained using the missing-data method in the spectral domain was reported to be better than those obtained using the conventional ASR in the cepstral domain. With an increase in the vocabulary size, however, the conventional ASR performs substantially better. Results using the missing value imputation methods have been reported on a larger vocabulary previously (Raj et al., 2000). Their method uses a binary mask and therefore is subject to the same limitations stated previously.

The rest of the paper is organized as follows. Section 2 provides an overview of the proposed systems. We then describe the binaural front-end for both the conventional and missing-data recognizers in Section 3. The section additionally provides the estimation details of ideal binary and ratio T-F masks. The conventional and missing-data recognition methods are reviewed in Section 4. The recognizers are tested on two different task domains with different vocabulary sizes. Section 5 discusses the two tasks and presents the evaluation results of the recognizers along with a comparison of their relative performance. Finally, conclusion and future work are given in Section 6.

2 System Overview

In this study, we analyze two strategies for robust speech recognition: 1) missing-data recognition and 2) a system that combines speech enhancement with a conventional ASR. The performance is examined at various SNR conditions and for two vocabulary sizes. Figure 1 shows the architecture of the two different processing strategies.

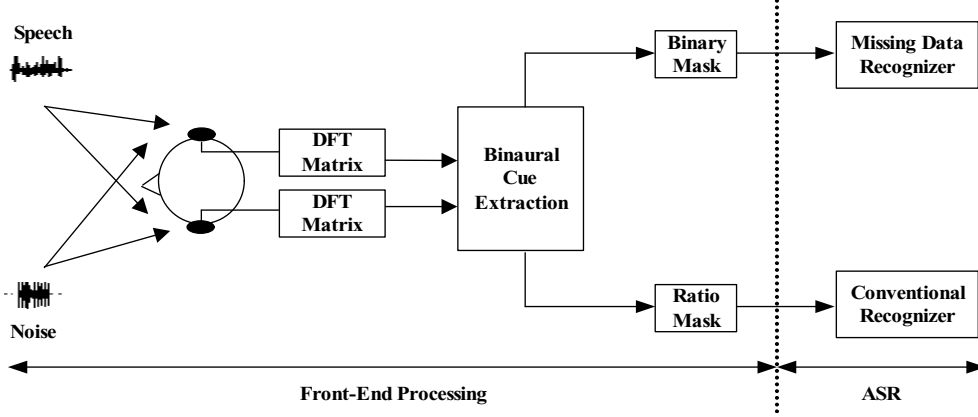


Fig. 1. Architecture of the two robust speech recognition strategies with binaural preprocessing: The missing-data recognizer and the conventional ASR. Left and right ear signals are obtained by filtering with HRTFs. A short-time Fourier analysis is applied to the signals, resulting in a time-frequency decomposition. ITD and IID are computed in each T-F unit. The missing-data recognizer works with a binary mask. A ratio mask is used as a speech enhancement strategy and is fed to the conventional recognizer.

The input to both systems is a binaural mixture of speech and interference presented at different, but fixed, locations. The binaural signals are obtained by filtering monaural signals with measured head-related transfer functions (HRTFs) corresponding to the direction of sound incidence. The responses to multiple sources are added at each ear. The HRTF measurements consist of left/right responses of the KEMAR manikin from a distance of 1.4 m in the horizontal plane, resulting in 128 point impulse responses at a sampling rate of 44.1 kHz (Gardner and Martin, 1994). HRTFs provide location-dependent ITD and IID which can be extracted independently in each T-F unit. The T-F resolution is 20 ms time frames with a 10 ms frame shift, and 512 DFT coefficients. Frames are extracted by applying a running Hamming window to the signal.

The missing-data speech recognizer operates in the log-spectral domain by using the knowledge about the reliability of spectral units at each time frame of the noisy speech input. Thus a binary mask that informs the recognizer of which T-F units are dominated by speech energy is required. A 64-channel auditory filterbank was used previously as a front-end for the missing-data recognizer (Cooke et al., 2001). We have chosen a DFT representation for the missing-data recognizer in order to be consistent with the conventional recognizer. A comparison between the DFT representation and the auditory

filterbank representation has shown that the difference in recognition performance is statistically insignificant. Statistics based on mixtures of multiple speech sources show that there exists a systematic correlation between the *a priori* energy ratio and the estimated ITD/IID values, resulting in a characteristic clustering across frequencies (Roman et al., 2003). To estimate the ideal binary mask we extend the non-parametric classification method of Roman et al. (2003) in the joint ITD/IID feature space independently for each frequency bin. The frequency decomposition used by Roman et al. (2003) was generated by a gammatone filterbank. This classification results in binary Bayesian decision rules that determine whether speech is stronger than interference in individual T-F units (energy ratio greater than 0.5). The system of Roman et al. (2003) was chosen because of the excellent match between their estimated binary mask and the ideal binary mask.

The conventional approach to robust speech recognition involves preprocessing of the corrupted speech by speech enhancement algorithms. This allows for the subsequent usage of decorrelating transformations (cepstral transformation, linear discriminant analysis) and temporal processing methods (delta features, RASTA filtering) on enhanced spectral features (Shire, 2000). In this study we use cepstral and delta features which are known to provide improved recognition accuracy. To enhance noisy speech we estimate an ideal ratio T-F mask. The statistics described above show that the estimated ITD and IID have a functional relationship with the *a priori* energy ratio. We employ this relationship in a non-parametric fashion to estimate the ideal ratio mask. Finally, to decode using the conventional ASR, MFCCs are computed from speech reconstructed after masking the corrupted signal by the estimated ratio mask.

3 A LOCALIZATION BASED FRONT-END FOR ASR

When speech and additive noise are orthogonal, the linear MMSE filter is the Wiener filter (Van Trees, 1968). With a frame-based processing, the MMSE filter corresponds to the ratio of *a priori* speech eigen values to the sum of *a priori* eigen values of speech and noise (Van Trees, 1968). Under asymptotic conditions, this corresponds to the frame-based Wiener filter (McAulay and Malpass, 1980; Van Trees, 1968). Ephraim and Malah (1984) have additionally shown that the optimal MMSE estimate of speech spectral amplitude in a local T-F unit is strongly related to the *a priori* SNR. To estimate the speech in a local T-F unit, we approximate the frame-based filter with an ideal ratio mask defined using the *a priori* energy ratio $R(\omega, t)$:

$$R(\omega, t) = \left[\frac{|S(\omega, t)|^2}{|S(\omega, t)|^2 + |N(\omega, t)|^2} \right], \quad (1)$$

where $S(\omega, t)$ and $N(\omega, t)$ are the target and noise spectral values at frequency ω and time t computed from the signal at the “better ear” - the ear with higher SNR. This is our computational goal for front-end processing with the conventional ASR.

In addition, an ideal binary mask assigns the label 1 to those T-F units whose energy ratio $R(\omega, t)$ exceeds 0.5 and assigns the label 0 otherwise. Such masks have been shown to generate high-quality reconstruction for a variety of signals and also provide an effective front-end for the missing-data recognition on a small vocabulary task (Cooke et al., 2001; Roman et al., 2003).

The objective of our front-end processing is to develop effective mechanisms for estimating both ideal binary and ratio masks. We propose an estimation method based on observed patterns for the binaural cues caused by the auditory interaction of multiple sources presented at different locations. Roman et al. (2003) have shown that for two sinusoidal signals, ITD and IID undergo systematic shifts as the energy ratio between the two sources changes. Moreover, statistics collected from real signals have shown similar patterns. In this case, training for each frequency bin is required since frequency-dependent combinations of ITD and IID arise naturally for a fixed spatial configuration. We employ the same training corpus as used by Roman et al. (2003) consisting of 10 speech signals from the TIMIT database (Garofolo et al., 1993). Five sentences correspond to the target location set and the rest belong to the interference location set. Binaural signals are obtained by convolving with KEMAR HRTFs as described in Section 2. This dataset is different from the databases used in training the ASRs.

The ITD/IID estimates are computed independently in each T-F unit based on the spectral ratio at the left and right ears:

$$(ITD, IID)(\omega, t) = \left[-\frac{1}{\omega} A \left(\frac{X_L(\omega, t)}{X_R(\omega, t)} \right), \frac{|X_L(\omega, t)|}{|X_R(\omega, t)|} \right], \quad (2)$$

where $X_L(\omega, t)$ and $X_R(\omega, t)$ are the left and right ear spectral values of the noisy speech at frequency ω and time t and $A(re^{j\phi}) = \phi$, $-\pi < \phi \leq \pi$. Note that at high frequencies, the phase is ambiguous corresponding to integer multiples of 2π . To disambiguate, we identify ITD in the range of $2\pi/\omega$ centered at zero delay.

Fig. 2 shows empirical results from the training corpus for a two-source configuration: target source in the median plane and interference at 30° . The scatter plot in Fig. 2A shows samples of ITD and R for a frequency bin at 1 kHz. Similarly, Fig. 2B shows the results that describe the variation of IID and R for a frequency bin at 3.4 kHz. The results are similar to those obtained by Roman et al. (2003), who use an auditory filterbank for frequency

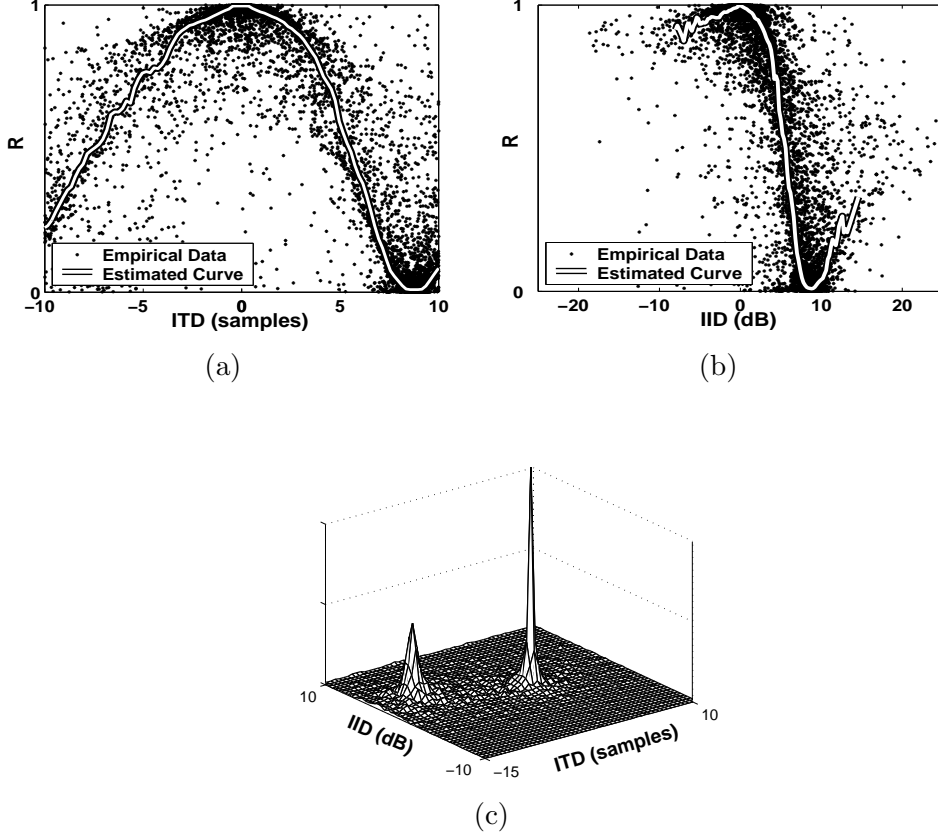


Fig. 2. Relationship between ITD/IID and the energy ratio R . Statistics are obtained with target in the median plane and interference on the right side at 30° . (a) The scatter plot for the distribution of R with respect to ITD for a frequency bin at 1 kHz. The solid white curve shows the mean curve fitted to the data. (b) Corresponding results for IID for a frequency bin at 3.4 kHz. (c) Histogram of ITD and IID samples for a frequency bin at 2 kHz.

decomposition. Note that the scatter plots exhibit a systematic shift of the estimated ITD and IID with respect to R . Moreover, a location-based clustering is observed in the joint ITD-IID space as shown in Fig. 2C. Each peak in the histogram corresponds to a distinct active source. Therefore, to estimate the ideal binary mask we employ non-parametric classification in the joint ITD-IID feature space as used by Roman et al. (2003). There are two hypotheses for the binary decision: H_1 - target is stronger or $R \geq 0.5$ and H_2 - interference is stronger or $R < 0.5$. Classification is obtained using the maximum *a posteriori* (MAP) decision rule: $p(H_1)p(x|H_1) > p(H_2)p(x|H_2)$ where x is (\hat{ITD}, \hat{IID}) (ω, t) feature vector. The prior probabilities, $p(H_i)$ are computed as the ratio of the number of samples in each class to the total number of samples. The conditional probabilities, $p(x|H_i)$ are estimated from the training

data using the kernel density estimation method (Roman et al., 2003).

In order to estimate the ideal ratio mask, we use the same training data. It is well known that ITD is salient at low frequencies while IID becomes more prominent at higher frequencies (Blauert, 1997). ITD exhibits different patterns across frequency bins as seen in the number of modes that characterizes the distribution of the samples (Roman et al., 2003). Hence, there is no unique parametric curve for all frequencies. Moreover, in the absence of evidence for a parametric estimate to provide better recognition results, a mean curve is fitted to the distribution of ITD. This is our estimate of the ideal ratio mask below 3 kHz. For higher frequencies, we utilize the information provided by the IID cues and use the same method to estimate the energy ratio. For improved results, we remove the outliers outside of 0.2 distance from the median. The resulting mean curves are shown in Fig. 2A (ITD) and Fig. 2B (IID). Thus, for given $\hat{ITD}(\omega, t)$ and $\hat{IID}(\omega, t)$, the estimated energy ratio $\hat{R}(\omega, t)$ is the corresponding value on the mean curve.

4 RECOGNITION STRATEGIES

We evaluate the binaural segregation system described in Section 3 as the front-end for robust ASR using two different recognizers. Conventional ASR uses MFCCs as the parameterization of observed speech. MFCCs are computed from the segregated speech obtained by applying the ratio mask to the noisy input signal. The missing-data recognizer uses log-spectral energy as feature vectors. This recognizer requires information about which T-F regions are reliable and which are unreliable. Thus the binary mask, generated by the binaural system, is additionally fed to the missing-data recognizer. A HMM toolkit, HTK (Young et al., 2000) is used in training of both recognizers and for the testing with the conventional ASR. During testing with the missing-data recognizer, the decoder is modified to incorporate the missing-data methods.

4.1 The Conventional Speech Recognizer

We use the standard continuous density HMM based speech recognizer trained on clean speech to model each word in the vocabulary (Section 5). Observation densities are modeled as mixture of Gaussians with diagonal covariance. The input to this ASR is the estimated speech spectral energy $|\hat{S}(\omega, t)|^2$.

$$|\hat{S}(\omega, t)|^2 = |X(\omega, t)|^2 \cdot \hat{R}(\omega, t), \quad (3)$$

where $|X(\omega, t)|^2$ is the spectral energy of the noisy signal at the better ear (see Section 5). From the estimated speech spectra, we compute the MFCCs. MFCCs are chosen as feature vectors as they are most commonly used in state-of-the-art recognizers (Rabiner and Juang, 1999). 13 cepstral coefficients along with delta and acceleration coefficients are extracted each frame, including the 0th order cepstral coefficient C_0 as the energy term. Frames are extracted as described in Section 2. A first-order preemphasis coefficient of 0.97 is applied to the signal.

4.2 The Missing-data Speech Recognizer

The missing-data recognizer (Cooke et al., 2001) makes use of spectro-temporal redundancy in speech to recognize a noisy speech based on its speech dominant T-F units. Given an observed speech vector Y , the problem of word recognition is to maximize the posterior $P(W_i|Y)$, where W_i is a valid word sequence according to the grammar for the recognition task. When parts of Y are corrupted by additive noise, it can be partitioned into its reliable and unreliable constituents as Y_r and Y_u . One can then seek the Bayesian decision given the reliable constituents. In the marginalization method, the posterior probability using only the reliable constituents is computed by integrating over the unreliable ones (Cooke et al., 2001). In missing-data methods, recognition is typically performed using spectral energy as feature vectors. If Y represents the observed spectrum and sound sources are additive, the unreliable parts may be constrained as $0 \leq Y_u^2 \leq Y^2$. This bounded marginalization method is shown by Cooke et al. (2001) to have a better recognition score than the simple marginalization method, and is hence used in all our experiments employing the missing-data recognizer. We use mixture of Gaussians with diagonal covariance to model the observed speech features as suggested by Cooke et al. (2001). Feature vectors for the missing-data recognizer comprise of 512 DFT coefficients per frame as described in Section 2. Log compression is applied to the resulting energy spectrum of the signal. Delta and acceleration coefficients are not calculated for log-spectral energy features due to problems in their use with the marginalization method (Raj, 2000). To provide the missing-data recognizer with the reliable and unreliable T-F units during decoding, we use the estimated binary mask as described in Section 3.

5 EVALUATION RESULTS

To compare the effect of vocabulary size on the two recognition approaches outlined above, we choose two task domains. The first task is speaker-independent recognition of connected digits. This is the same task used in

the original study of Cooke et al. (2001). Thirteen (1-9, a silence, very short pause between words, zero and oh) word-level models are trained for both recognizers. All except the short pause model have 8 emitting states. The short pause model has a single emitting state, tied to the middle state of the silence model. The output distribution in each state is modeled as a mixture of 10 Gaussians, as suggested by Cooke et al. (2001). The grammar for this task allows for one or more repetitions of digits. All digits are equally probable. The TIDigits database’s male speaker data is used for both training and testing (Leonard, 1984). Specifically, the models are trained using 4235 utterances in the training set of this database. Testing is performed on a subset of the testing set consisting of 232 utterances from 3 speakers. All test speakers are different from the speakers in the training set. The signals in this database are sampled at 20kHz.

The second task is the speaker-independent recognition of command and control type phrases. Two hundred and eight (206 words, a silence and a short pause between words) word-level models are trained for both recognizers. This task allows us to increase the vocabulary size from thirteen to two hundred and eight, a natural progression in testing the effect of vocabulary size on the recognizers. All except the short pause model have 8 emitting states, whose output distribution is modeled as a mixture of 8 Gaussians. The short pause model has a single state. The number of Gaussians in the mixture is slightly lower compared to the digit recognition task due to the lack of adequate training data for some of the models. The grammar for this task assigns equal probability to all phrases in the database. For a given phrase, the word sequence is fixed. The digital data subset of the Apple Words and Phrases database is used for both training and testing (Cole et al., 1995). In particular, 1996 speakers with IDs 21 through 2604 are used for training. This corresponds to 63835 utterances. Data from 14 speakers with IDs 4 through 19 are used for testing. This corresponds to 454 utterances. The signals are sampled at 8kHz.

The two tasks also differ in perplexity. Perplexity is one indicator of difficulty of the recognition task along with vocabulary size (Rabiner and Juang, 1999). For the digit recognition task, the perplexity is 11.0. For the command and control task the perplexity is 3.05. For our task, we calculate the perplexity empirically from the word level lattice (Young et al., 2000). The lower perplexity for the second task is due to the use of a restrictive grammar for this task (Cole et al., 1995). To test the robustness of the two recognizers in the aforementioned tasks, noise is added at a range of SNRs from -5 dB to 10 dB in steps of 5 dB. Higher positive values of SNRs are not explored, as one of the recognizers saturates to ceiling performance at 10 dB. The noise source for both recognition tasks is the factory noise from the NOISEX corpus (Varga et al., 1992), which is also used by Cooke et al. (2001). The factory noise is chosen as it has energy in the formant regions, therefore posing challenging problems for recognition. It is also impulsive, making it difficult to estimate

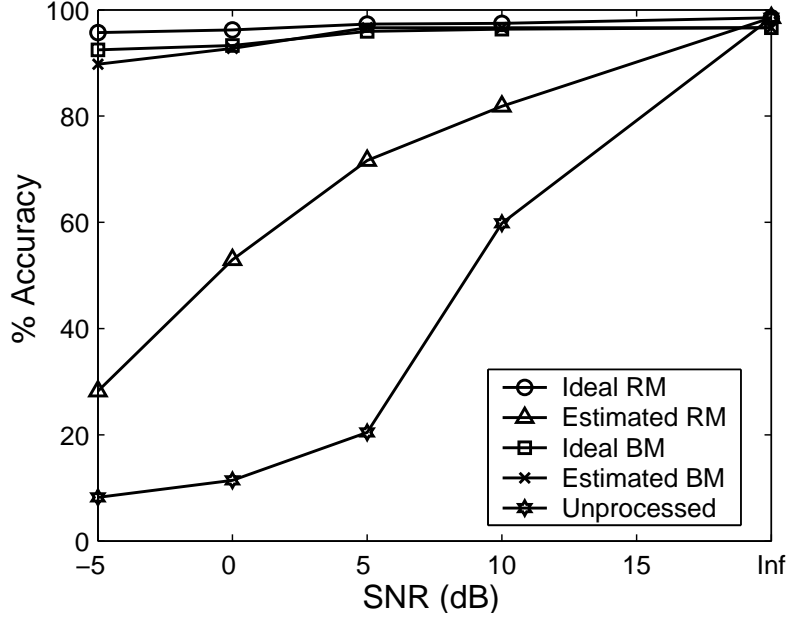


Fig. 3. Performance of conventional and missing-data recognizers on the digits recognition task. Ideal RM refers to the performance of the conventional ASR using the ideal ratio mask. Estimated RM refers to its performance when using the estimated ratio mask by the binaural front-end. Ideal BM refers to the performance of the missing-data ASR using the ideal binary mask. Estimated BM refers to the performance of the same when using the estimated binary mask by the binaural front-end. For comparison, the performance of the conventional ASR without the use of any front-end processing is also shown.

its spectrum using spectral subtraction methods (Cooke et al., 2001). In all our experiments, the target speech source is in the median plane and the noise source on the right side at 30° , making the left ear the better ear in terms of SNR (see Section 3).

Fig. 3 summarizes the performance of the two recognizers on the digit recognition task. Performance is measured in terms of word-level recognition accuracy under various SNR conditions. “Unprocessed” refers to the baseline performance of the conventional ASR, without the use of any front-end processing. The figure shows the recognition accuracy of the conventional ASR with the use of ideal and estimated ratio T-F masks (“Ideal RM” and “Estimated RM” respectively). This is compared to the accuracy of the missing-data recognizer, which uses ideal and estimated binary T-F masks (“Ideal BM” and “Estimated BM” respectively).

Fig. 3 shows the robust performance of the ideal ratio mask when used as

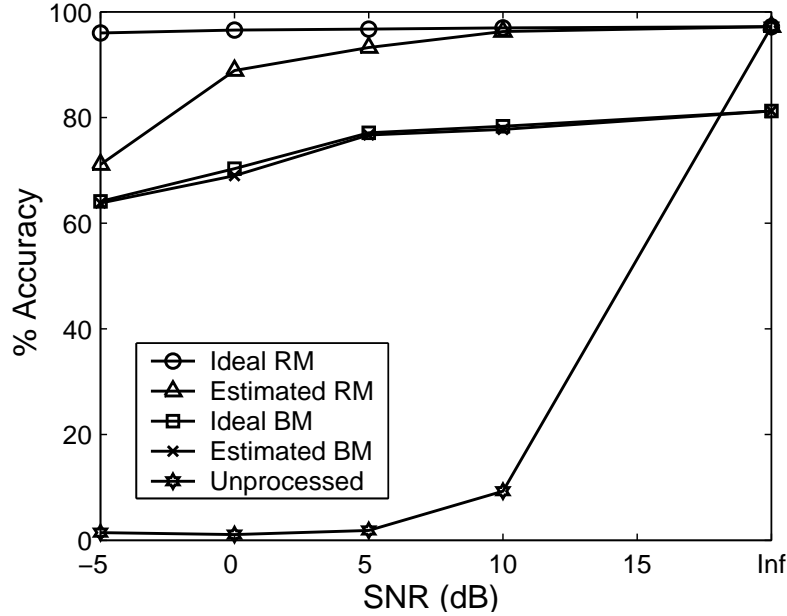


Fig. 4. Performance of conventional and missing-data recognizers on the command and control task. See Figure 3 caption for notations.

a front-end for conventional ASR. Only a minor performance degradation is observed even at -5 dB. The performance of the conventional ASR with the estimated ratio mask degrades much faster than with the ideal ratio mask. This indicates that the conventional ASR is sensitive to errors in our estimation of the ideal ratio mask. Observe that the performance with the use of the estimated ratio mask is still substantially better than that with no pre-processing across all SNR conditions. As reported by Cooke et al. (2001), the performance of the missing-data recognizer degrades very little with increasing amounts of noise added, indicating the adequacy of recognition using a binary mask for this task. Also, the performance with the estimated binary mask is close to that with the ideal binary mask, indicating the high quality of the front-end to estimate the ideal binary mask (see also Roman et al., 2003). Notice that, for this task, the performance of the missing-data recognizer is close to the performance of the conventional ASR with the ideal ratio mask.

Similarly, Fig. 4 summarizes the performance of the two recognizers on the task of recognition of command and control phrases. The relative performance of the two recognizers reverses with this increase in the vocabulary size. As in the digits recognition task, the performance of the conventional ASR using the ideal ratio mask is close to the ceiling performance. Additionally, its performance using the estimated ratio mask is close to the that with the ideal ratio mask, especially at $\text{SNR} > 0$ dB. The increased accuracy of the con-

ventional ASR using the estimated ratio mask compared to its performance on the digits recognition task is due to the lower perplexity of this task. Its performance now is substantially better than that of the missing-data recognizer using both ideal and estimated binary masks, particularly at $\text{SNR} \geq 0$ dB. Notice that the performance of the missing-data recognizer with the estimated binary mask is close to that with the ideal binary mask as in the digits recognition task, confirming the ability of the front-end to estimate the ideal binary mask accurately.

Lower accuracy values for the missing-data recognizer using both binary masks in Fig. 4 may be attributed to a number of reasons. It is known that the use of mixtures of Gaussians with diagonal covariance structure does not adequately represent the observed spectral vectors (Cooke et al., 2001) and this problem gets exacerbated with an increase in the vocabulary size. Thus, under clean speech conditions, the difference between the accuracy of conventional and missing-data recognizers increases with increase in vocabulary size (see also Raj et al., 2000). One could compute MFCCs from the speech resynthesized using the binary T-F masks, and use it for decoding. Under clean speech conditions, the missing-data recognizer would then have the same recognition accuracy as that of the conventional ASR. The performance though degrades rapidly with decreasing SNR (de Veth et al., 1999).

The use of binary masks does not compensate for amplitude distortions, because the mixture spectral values are used in recognition for those T-F units labeled 1. Could this be the reason for reduced performance in larger vocabulary recognition? To test the effect of this distortion, we replace the spectral vectors of the reliable T-F regions with their corresponding clean speech values, calculated *a priori*. The performance, at various SNRs, is summarized in Tables 1 and 2. “Distorted” refers to the performance of the missing-data recognizer on the mixture spectral values for all T-F units. “Undistorted” refers to its performance when the reliable T-F units contain clean speech values. In the unreliable units, we retain the spectral values of noisy speech. We use the ideal binary mask generated at each SNR to provide the reliability information for both conditions. Table 1 shows the effect of amplitude distortion on the digits recognition task. For this task, the effect of amplitude distortion is seen, as expected, to be minimal across all SNRs, since the recognition accuracy is already quite high. Table 2 shows the effect of amplitude distortion on the task of command and control phrases. Except at 0 dB SNR condition, only a small improvement is observed by eliminating the noise energy from the reliable T-F units. Hence, the degradation to the overall performance of the missing-data recognizer caused by this amplitude distortion is statistically insignificant at the range of SNRs considered here. When using the ideal binary mask generated at each SNR directly on clean speech, we observe a degradation in performance. This may be attributed to the use of energy bounds for the unreliable units in the marginalization method.

Table 1

Effect of amplitude distortion in the reliable T-F regions on recognition accuracy (%) of the missing-data recognizer for the digits recognition task.

<i>Amplitude</i>	<i>SNR (dB)</i>			
	-5	0	5	10
Distorted	92.47	93.28	95.97	96.37
Undistorted	93.68	94.89	95.30	95.43

Table 2

Effect of amplitude distortion in the reliable T-F regions on recognition accuracy (%) of the missing-data recognizer for the command and control task.

<i>Amplitude</i>	<i>SNR (dB)</i>			
	-5	0	5	10
Distorted	64.14	70.33	77.11	78.35
Undistorted	65.86	73.85	77.69	80.21

Comparing Figures 3 and 4, we can see that the performance curve for the missing-data recognizer is steeper on the second task compared to the first task. Note that this behavior is opposite to that of the conventional ASR. While the conventional ASR performs better on the second task by utilizing the lower perplexity of the language model, the missing-data recognizer is unable to do so. This may be caused by the inability of the missing-data recognizer to represent all the speech models adequately. The log-spectral representation may have a limited expressibility in terms of distinct words that can be uniquely represented. The TIDigits database has a small vocabulary. The Applewords database with a larger vocabulary creates many more competing models during decoding. Thus, within the same T-F grid, an increased number of words need to be discriminated. With the use of a binary mask, only a small portion of the total T-F acoustic model space is utilized during recognition. This makes it difficult for the missing-data recognizer to differentiate between competing hypothesis. Fig. 5 shows the effect of using the same binary T-F mask on two signals. Fig. 5(a) shows the spectrogram of the word “Billy” and Fig. 5(b) shows the spectrogram of the word “Delete”. Fig. 5(c) shows a typical ideal binary T-F mask generated at low SNR. The reliable units in this mask are white and the unreliable black. This binary mask is applied to the spectrograms in Fig. 5(a) and Fig. 5(b) and the resulting spectrograms with only reliable T-F units are shown in Fig. 5(d) and Fig. 5(e), respectively. Notice that the reliable regions of the two spectrograms are very

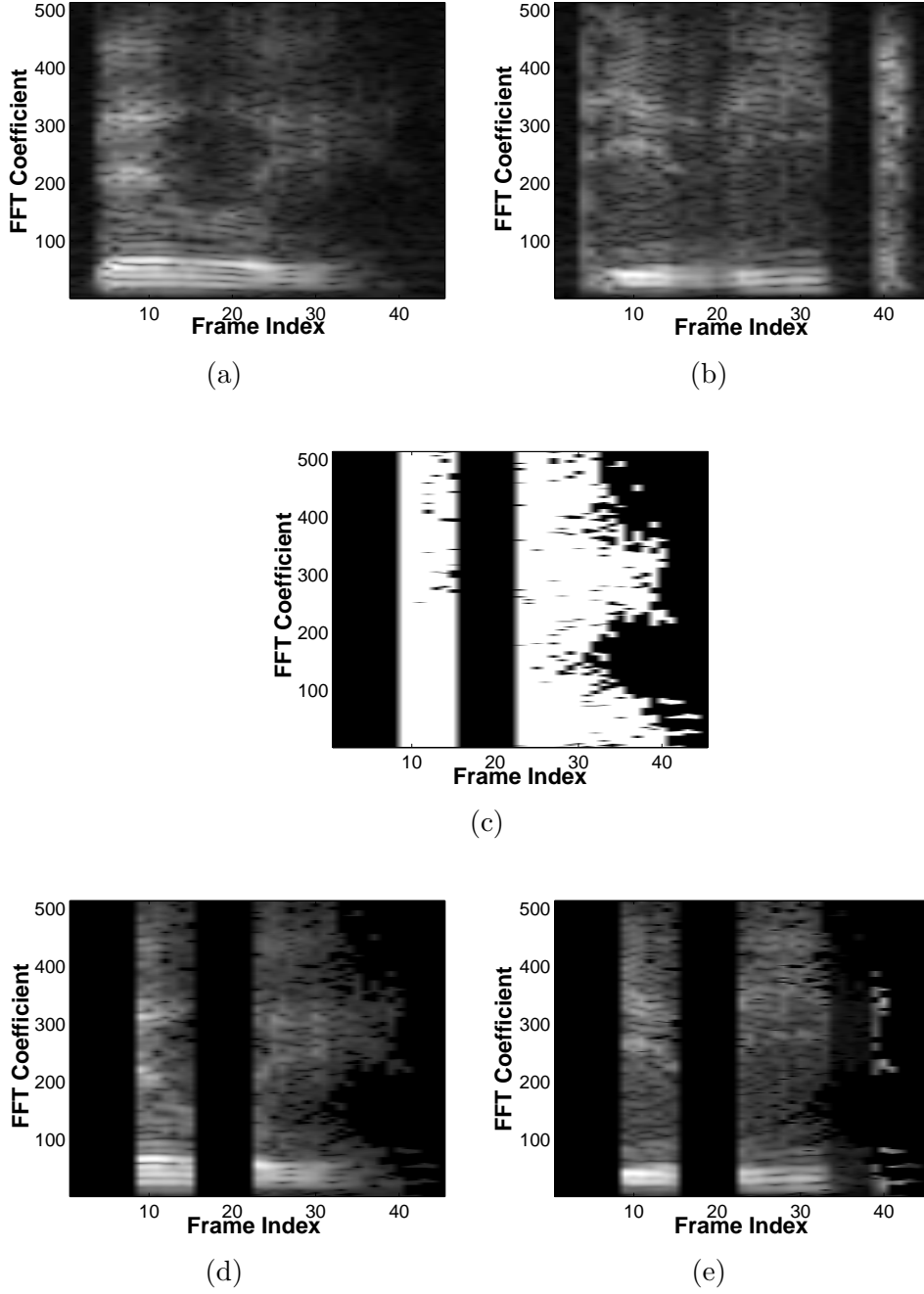


Fig. 5. An illustration of similarity of reliable regions. (a) The spectrogram of the word “Billy”. (b) The spectrogram of the word “Delete”. (c) An ideal-binary T-F mask. Reliable T-F units are marked white and unreliable black. (d) The spectrogram obtained from (a) by applying the ideal mask in (c). (e) The spectrogram obtained from (b) by the same ideal masking as in (d).

similar. In the absence of information in the unreliable regions, it is difficult for the recognizer to distinguish between the two words. Indeed the recognizer frequently substitutes one word with the other. The bounded marginalization method treats the information in the unreliable regions only as counter-evidence for recognition of certain models (Cunningham and Cooke, 1999). Hence, the missing-data recognizer faces increased acoustic complexity during decoding.

6 DISCUSSION

The advantage of the missing-data recognizer is that it imposes a lesser demand on the speech enhancement front-end than the conventional ASR. Only knowledge of reliable T-F units of noisy speech, or an ideal binary mask, is required from the front-end. Moreover, Roman et al. (2003) have shown that the performance of the missing-data recognizer degrades gradually with increasing deviation from the ideal binary mask. The binaural system employed here is able to estimate this mask accurately. Hence, we achieve performance close to the ceiling performance of missing-data recognition. Conventional ASR on the other hand, requires full-band speech enhancement by a front-end processor. In this study, we have employed a ratio T-F mask as a front-end for the conventional ASR, which is estimated using statistics of ITD and IID. Estimation of the ideal ratio mask is less robust than the estimation of the ideal binary mask. The conventional ASR is very sensitive to errors of such front-end processing (Barker et al., 2004; Raj, 2000). As a result, the performance of the missing-data recognizer on the small vocabulary task is better than that of the conventional ASR.

The marginalization method for missing-data recognition is the optimal spectral domain recognition strategy provided that the missing T-F units can be ignored for classification (Little and Rubin, 1987). The missing-data recognizer assumes that the unreliable units carry redundant information for speech recognition. This, however, is not always true. For a small vocabulary task, the unreliable units may be safely marginalized for good recognition results. When vocabulary size increases, the acoustic model space becomes densely populated. Under such conditions, good recognition results may not be obtained by completely ignoring the missing T-F units. This may be caused by the inability to represent all the acoustic models adequately using only a small number of reliable T-F units. On the other hand, the ratio T-F mask attempts to recover the speech in the unreliable T-F units for use in recognition. Additionally, under clean speech conditions, recognition accuracy using spectral features is inferior to using cepstral features. The cepstral transformation retains the envelope of speech while removing its excitation source (Rabiner and Juang, 1999). The speech envelope contains most relevant information for

recognition. In addition, cepstral features are used for their quasi-orthogonal properties (Shire, 2000). Hence, advantage of the conventional ASR shows when vocabulary size increases. Our experiments also suggest that the sensitivity of the conventional ASR to speech enhancement errors may be mitigated by employing better language models. Better language models though seem less effective in overcoming the data paucity limitation of the missing-data recognizer.

Raj et al. (2000) have previously reported that conventional ASR with reconstructed missing T-F regions outperforms the missing-data recognizer when tested on the Resource Management database (Price et al., 1988). The missing or unreliable T-F units were reconstructed either using speech clusters or based on their correlations with reliable regions. The speech clusters and the knowledge of correlations between reliable and unreliable T-F units are obtained from the training portion of the Resource Management database. Unlike their system, our estimation of the ideal ratio mask is independent of the signals used in the training and testing of the speech recognizers. Hence, it is applicable even when samples of clean speech are unavailable. Additionally, the accuracy and computational complexity of our ratio mask based system are not dependent on the nature and size of the vocabulary.

Although our estimated T-F ratio mask provides promising results, other approaches for the estimation of this mask could also be explored; should a parametric curve be suspected, the parameters could be optimized to minimize recognition errors. Future work will also extend to large vocabulary tasks and explore the robustness of the binaural front-end to changes in location and number of noise sources.

To summarize, we have proposed a ratio T-F mask, estimated using a binaural processor, as a front-end for conventional ASR. At two different vocabulary sizes, the use of this mask results in significant improvement in recognition accuracy at various SNRs when compared to the baseline performance. On the larger vocabulary task, the performance of the proposed ASR is substantially better than that of the missing-data recognizer. Our study suggests that optimal preprocessing strategies for robust speech recognition may depend on the vocabulary size of the task. For small vocabulary applications, computation of the ideal T-F binary mask may be desirable, whereas a ratio mask may provide an improved performance with increased vocabulary sizes.

Acknowledgements

This research was supported in part by an AFOSR grant (FA9550-04-1-0117) and an NSF grant (IIS-0081058). We thank M. Cooke for discussion and assistance in implementing the missing data recognizer.

References

- Barker, J., Cooke, M., Ellis, D. P. W., 2004. Decoding speech in the presence of other sources. *Speech Communication*, in press.
- Blauert, J., 1997. *Spatial Hearing - The psychophysics of human sound localization*. The MIT Press, Cambridge, MA.
- Boll, S. F., 1979. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. on Acoustics, Speech, and Signal Processing ASSP-27* (2), 113–120.
- Brandstein, M., Ward, D. (Eds.), 2001. *Microphone arrays: Signal processing techniques and applications*. Springer, Berlin, Germany.
- Cardoso, J. F., 1998. Blind signal separation: Statistical principles. *Proc. IEEE* 86 (10), 2009–2025.
- Cole, R., Noel, M., Lander, T., Durham, T., 1995. New telephone speech corpora at CSLU. In: *Proc. EUROSPEECH '95*. pp. 821–824.
- Cooke, M., Green, P., Josifovski, L., Vizinho, A., 2001. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication* 34, 267–285.
- Cunningham, S., Cooke, M., 1999. The role of evidence and counter-evidence in speech perception. In: *Proc. International Congress on Phonetic Sciences '99*. pp. 215–218.
- Davis, S. B., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. on Acoustics, Speech, and Signal Processing ASSP-28* (4), 357–366.
- de Veth, J., de Wet, F., Cranen, B., Boves, L., 1999. Missing feature theory in ASR: Make sure you miss the right type of features. In: *Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions '99*. pp. 231–234.
- Droppo, J., Acero, A., Deng, L., 2002. A nonlinear observation model for removing noise from corrupted speech log mel-spectral energies. In: *Proc. International Conference on Spoken Language Processing '02*. pp. 1569–1572.
- Ehlers, F., Schuster, H. G., 1997. Blind separation of convolutive mixtures and an application in automatic speech recognition in a noisy environment. *IEEE Trans. on Signal Processing* 45 (10), 2608–2612.
- Ephraim, Y., 1992. A Bayesian estimation approach for speech enhancement using hidden Markov models. *IEEE Trans. on Signal Processing* 40 (4), 725–735.
- Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. on Acoustics, Speech, and Signal Processing ASSP-32* (6), 1109–1121.
- Gales, M. J. F., Young, S. J., 1996. Robust continuous speech recognition using parallel model combination. *IEEE Trans. on Speech, and Audio Processing* 4 (5), 352–359.
- Gardner, W. G., Martin, K. D., 1994. HRTF measurements of a KEMAR

- dummy-head microphone. Technical Report #280, MIT Media Lab Perceptual Computing Group.
- Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallet, D., Dahlgren, N., 1993. Darpa TIMIT acoustic-phonetic continuous speech corpus. Technical Report NISTIR 4930, National Institute of Standards and Technology, Gaithersburg, MD.
- Glotin, H., Berthommier, F., Tessier, E., 1999. A CASA-labelling model using the localisation cue for robust cocktail-party speech recognition. In: Proc. Eurospeech'99. pp. 2351–2354.
- Gong, Y., 1995. Speech recognition in noisy environments: A survey. *Speech Communication* 16, 261–291.
- Hughes, T. B., Kim, H. S., DiBase, J. H., Silverman, H. F., 1999. Performance of an HMM speech recognizer using a real-time tracking microphone array as input. *IEEE Trans. on Speech, and Audio Processing* 7 (3), 346–349.
- Leonard, R. G., 1984. A database for speaker-independent digit recognition. In: Proc. ICASSP '84. pp. 111–114.
- Lippmann, R. P., 1997. Speech recognition by machines and humans. *Speech Communication* 22, 1–15.
- Little, R. J. A., Rubin, D. B., 1987. Statistical analysis with missing data. John Wiley and Sons, New York, NY.
- McAulay, R., Malpass, M. L., 1980. Speech enhancement using a soft-decision noise supression filter. *IEEE Trans. on Acoustics, Speech, and Signal Processing ASSP-28* (2), 137–145.
- Palomaki, K. J., Brown, G. J., Wang, D. L., 2004. A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation. *Speech Communication*, in press.
- Price, P., Fisher, W. M., Bernstein, J., Pallett, D. S., 1988. The DARPA 1000 word Resource Management database for continuous speech recognition. In: Proc. ICASSP '88. pp. 651–654.
- Rabiner, L. R., Juang, B. H., 1999. Fundamentals of speech recognition, 2nd Edition. Prentice-Hall, Englewood Cliffs, NJ.
- Raj, B., 2000. Reconstruction of incomplete spectrograms for robust speech recognition. Ph.D. thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University.
- Raj, B., Seltzer, M. L., Stern, R. M., 2000. Reconstruction of damaged spectrographic feautres for robust speech recognition. In: Proc. International Conference on Spoken Language Processing '00. pp. 1491–1494.
- Roman, N., Wang, D. L., Brown, G. J., 2003. Speech segregation based on sound localization. *J. Acoust. Soc. Am.* 114, 2236–2252.
- Rosenthal, D. F., Okuno, H. G. (Eds.), 1998. Computational auditory scene analysis. Lawrence Erlbaum Associates, Mahwah, NJ.
- Shire, M. L., 2000. Discriminant training of front-end and acoustic modeling stages to heterogeneous acoustic environments for multi-stream automatic speech recognition. Ph.D. thesis, University of California, Berkeley.
- Tessier, E., Berthommier, F., Glotin, H., Choi, S., 1999. A CASA front-end

- using the localisation cue for segregation and then cocktail-party speech recognition. In: Proc. IEEE International Conference on Speech Processing. pp. 97–102.
- Van Trees, H. L., 1968. Detection, Estimation, and Modulation Theory, Part I. John Wiley and Sons, New York, NY.
- Varga, A. P., Moore, R. K., 1990. Hidden Markov model decomposition of speech and noise. In: Proc. ICASSP '90. pp. 845–848.
- Varga, A. P., Steeneken, H. J. M., Tomlinson, M., Jones, D., 1992. The NOISEX-92 study on the effect of additive noise on automatic speech recognition. Technical Report, Speech Research Unit, Defense Research Agency, Malvern, UK.
- Wang, D. L., Brown, G. J., 1999. Separation of speech from interfering sounds based on oscillatory correlation. IEEE Trans. on Neural Networks 10 (3), 684–697.
- Young, S., Kershaw, D., Odell, J., Valtchev, V., Woodland, P., 2000. The HTK Book (for HTK Version 3.0). Microsoft Corporation.